INFO-H600 Computing Foundations of Data Sciences
Project: Drivers and Zones

Data

In this project, you will be asked to implement a data processing pipeline for computing driving habits of drivers.

Information about the position of each driver is produced from the GPS unit of mobile phones. You need to read the data from the provided CSV file `drivers.csv`. Each row of the CSV refers to one GPS measurement. The schema of the drivers data is the following:

- driver (String), the unique identifier of each driver
- Timestamp (Timestamp), the event creation time
- latitude (Double), the latitude geographic coordinate as received from GPS sensors
- longitude (Double), the longitude geographic coordinate as received from GPS sensors

Information about the zones is contained in the provided JSON file `zones.json`. Each zone is described by a polygon and a unique identifier. A polygon is described by a counterclockwise sequence of points (latitude - longitude pairs). The schema of the zones data is the following:

- id_zone (Long), the unique identifier of each zone
- polygon (Array of points), sequence of points (latitude - longitude pairs)

A polygon cannot have holes. Note that there is no guarantee that all driver positions belong within a polygon; this might be the result of a driver leaving the city or GPS jittering.

Questions

For the following exercises you should use Dask or Spark as your main data processing tool. You should write a report explaining and discussing your approach.

Ex. 1. Given the positions of *drivers* and a list of different *zones*, geographic areas, you will have to find the *ten zones that are visited by the most drivers*. Discuss how parallelization works in your analysis. Suggest and critique alternate parallelization strategies.

Ex. 2. Given this data, perform some other interesting data analysis. What additional knowledge would you like to extract from this data? As before, comment on the parallelization strategy.

Instructions

This project is to be made in groups of four persons. You are asked to form the groups via the activity "Groups for Project" on UV.

You should submit a *single* zip file containing the report and any documented python code you want to share.

The single file has to be uploaded on UV by **January 19, 2024**. There should be one submission per group.